Evaluation of an Expert System Producing Geometric Solids as Output

Case H. Ketting, M.D., Mary M. Austin-Seymour, M.D., Ira J. Kalet, Ph.D.
Jonathan P. Jacky, Ph.D., Sharon E. Kromhout-Schiro, Ph.D.*
Sharon M. Hummel, R.T(T), C.M.D., Jonathan M. Unger, M.S.†
Radiation Oncology Dept., University of Washington, Seattle, WA 98195

Lawrence M. Fagan, M.D., Ph.D.
Section on Medical Informatics, Department of Medicine
Stanford University, Stanford, CA 94305

Abstract

This paper reports the evaluation of an expert system whose output is a three-dimensional geometric solid. Evaluating such an output emphasizes the problems of establishing a comparison standard, and of identifying and classifying deviations from that standard. Our evaluation design used a panel of physicians for the first task and a separate panel of expert judges for the second. We found that multi-parameter or multi-dimensional expert system outputs, such as this one, may result in lower overall performance scores and increased variation in acceptability to different physicians. We surmise that these effects are a consequence of the higher number of factors which may be deemed unacceptable. The effects appear, however, to be equal for computer and human output. This evaluation design is thus applicable to other expert systems producing similarly complex output.

PROBLEM DEFINITION

Most medical expert systems produce discrete symbolic output, such as a diagnosis or a specific therapeutic recommendation. Such outputs are easily interpreted, and in evaluation of the system, determination of their correctness is straightforward. Typically, comparison of the output to an objective standard, such as a pathologic finding or textbook definition, results in a direct assessment of agreement or disagreement. Where no such gold standard exists, a practical substitute is made by comparing the expert system output to that of several human experts.

In contrast, the expert system described in this article produces as its output a Planning Target Volume (PTV), an irregular geometric solid used in radiation therapy treatment planning.[1] Since this volume is represented as a collection of polygon vertices, the evaluation of its correctness, even by comparison to a gold standard, is non-trivial. The vertices, centroids, and calculated volumes of an acceptable PTV and the gold standard would be unlikely to exactly agree. Furthermore, approximate numeric agreement of these parameters does not necessarily indicate clinical validity, nor do other common measures such as degree of overlap. For example, contours showing 99% agreement by some mathematical measure might differ strongly in clincal acceptability if the 1% difference caused one contour to overlap the spinal cord or some other radiaiton-sensitive structure. To circumvent these difficulties, we used expert evaluators to determine acceptability of PTVs. This obviated the need for a mathematical comparison methodology.

It is expected that, with the proliferation of expert systems, outputs of this complex and multidimensional nature will become more common. Our results emphasize the importance of careful evaluation design, as these outputs can significantly affect performance scores. Furthermore, we observed the utility of such evaluations for pointing out errors prevalent in clinical practice.

THE PTV TOOL

As part of the advance toward true three dimensional radiation therapy treatments, the International Commission on Radiation Units recently

*current address: Department of Biomedical Engineering, University of North Carolina, Chapel Hill, North Carolina 27599

† current address: RSA, Inc., 22 Terry Ave., Burlington, MA 01803

0195-4210/95/$5.00

defined an entity known as the planning target volume (PTV).[1] While a physician should ideally be able to outline a tumor volume on a patient's treatment planning CT scan and then irradiate only that volume, certain realistic constraints make this impossible. During a radiation treament, the tumor may move relative to its location on the CT scan due to physiologic causes such as respiration or gastrointestinal peristalsis. Additionally, since treatments are given daily over a period of weeks, further error is introduced from variations in daily treatment set-up. Consequently, in order to ensure treating the entire tumor volume to full dose, an expansion of the tumor volume is defined which takes into account these intra- and inter-treatment uncertainties. This larger volume, illustrated by the dashed line in Figure 1, is the planning target volume.

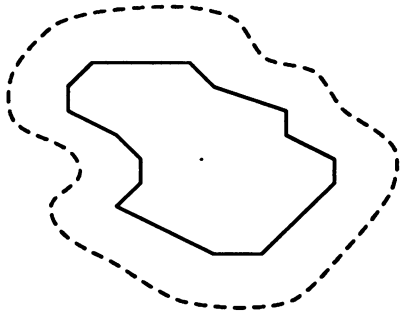

Figure 1: Tumor volume in the x,y (CT) plane showing adequate surrounding PTV.

The Planning Target Volume Tool (PTVT) generates a PTV using a knowledge base, inference engine, and volume expansion algorithm. It relies on the provided tumor volume as a basis for expansion, and uses information about the tumor location, histology, stage, and patient immobilization to derive the parameters of the expansion. Currently, the knowledge base covers nasopharyngeal carcinoma and non-small cell lung cancers.

EVALUATION METHODOLOGY

Typically, expert system evaluation involves direct comparison of the computer output with that of human experts. However, when the output is a geometric solid, as in this case, comparison proves problematic. While quantitative and qualitative statements may be made about the differences between the output solids, one must still define which of these differences are crucial, and for those that are, where the bounds of acceptability lie. Results of the evaluation process will depend heavily upon who makes these definitions and how they are applied.

In order to deal with these issues, our experimental design for tool evaluation relies on two separate groups of physician experts. One group, designated generators, produces reference PTVs for comparison with the PTVT output. The second group, designated evaluators, serves to establish criteria for judging the PTVs and to perform the judging process according to their criteria.

The evaluation methodology is a modification of the Turing test, similar to that described by Wyatt and Spiegelhalter,[2] in which a panel of experts evaluated the output of a computer system against a gold standard. For our evaluation, the comparison standard was the output of the physician generators. The PTVs were not identified by source, so the evaluators were unaware at any time whether they were judging the output of a human or a computer. Thus, if their scoring showed no significant difference between human and computer-generated PTVs, the PTV Tool could be said to have passed a version of the Turing test.

Notably, this approach circumvents the problem of identifying a gold standard by generating its own internal standard in the evaluation criteria used by the physician evaluators. It avoids the problem of defining a clinically meaningful comparison algorithm for irregular geometric solids by using the evaluators to do the comparison, then statistically testing their scoring of the PTVs.

Because the PTVT is expected to fall within the range of variation of the human generators, the null hypothesis states that the human and computer generators performed equally well. The alternate hypothesis states that significant performance differences exist.

EVALUATION PROCESS

All participating physicians were experienced in three-dimensional treatment planning and were selected from three different institutions. The four physician generators were presented with the test patient data shown in Table 1. This was accompanied by a normal CT scan with a simulated tumor volume drawn on the appropriate slices. (Pathologic CT scans were not used, as early trials showed disagreement to arise over the accuracy of the drawn tumor volumes. Such disagreement was thought to affect the resulting PTV.) Using this information, they then generated planning target volumes. The PTVT generated planning target

volumes for comparison, using the same data provided to the physician generators.

Table 1: Test Patient Data: Cases are labeled NP for nasopharyngeal, L for lung, SqCa is squamous cell carcinoma, and all the lung cases are adenocarcinomas.

Label	Description	Stage	Immob. device
L1	R hilar	T4N2	none
L2	L hilar	T4N2	none
L3	L upper lobe	T2N0	none
L4	L upper lobe	T3N0[a]	none
L9	R lower lobe	T2N2	none
NP1	SqCa	T4N0	mask
NP1a	SqCa	T4N0	none
NP3	SqCa	T2N0	mask
NP4	SqCa	T4N2	mask

[a] with chest wall fixation

In order to establish a sense of the process, the three evaluators also drew PTVs independently for two of the test cases. They compared and discussed their results, using this as a basis for establishing judging criteria. They were then provided with the output of the generators, both human and computer.

Each evaluator, using their mutual criteria, independently designated each PTV as acceptable or unacceptable. When a PTV was deemed unacceptable, the evaluator was required to state a reason.

The evaluation produced 135 data points, representing the scoring (acceptable vs. unacceptable) given by each evaluator for the five PTVs (four human and one computer) generated on each of the nine patients. Logistic regression analysis was used to determine the effects of patient, PTV generator, and evaluator on the score. This method models the logarithm of the odds as a linear function of the independent variables,

$$\log(\frac{p}{1-p}) = b_0 + b_1 \cdot pred_1 + b_2 \cdot pred_2 + \ldots \quad (1)$$

where p is the probability of any particular PTV being scored as acceptable and the $pred_i$ are the independent predictor variables to be tested for effect on this probability.

The coefficients, $b_0 \ldots b_n$, can be tested for significance using the score test, which has a χ^2 distribution. Coefficients which reach significance imply that the associated predictor variable (generator, evaluator or patient) exerts a statistically significant effect on the measured variable (pass rate).

An additional qualitative analysis was done on the reasons given for rejecting a PTV. By classifying these reasons into categories, it was possible to determine if the nature of the errors made by the PTVT was different from those made by human generators.

RESULTS

The pass rates as a function of patient, generator, and evaluator are listed in Table 2. Logistic regression analysis showed the PTV pass rate to have a strong dependence upon both evaluator and patient ($p < 0.05$).

Table 2: Human/Computer Pass Rate (%) by Patient and Evaluator. The columns show separately human (H) and computer (C) generated data. The average at the bottom is the weighted combination of human and computer generated data. The labels correspond to Table 1.

	Eval 1		Eval 2		Eval 3		All Evals	
Label	H	C	H	C	H	C	H	C
L1	75	100	0	0	75	100	50	67
L2	75	100	0	0	75	100	50	67
L3	25	0	0	0	75	100	33	33
L4	0	0	50	0	75	0	42	0
L9	100	0	50	0	75	0	58	0
NP1	50	0	0	0	25	100	25	33
NP1a	0	0	0	0	50	0	17	0
NP3	50	100	0	0	25	100	25	67
NP4	0	0	0	0	50	100	17	33
Tot.	42	33	11	0	58	67	36	33
Avg.	40		9		60		35	

The first finding is clearly evident from a perusal of Table 2. The pass rate for PTVs reviewed by Evaluator 2 was less than 9%, while the other two evaluators produced pass rates of 40% and 60%. The statistical analysis does not, of course, give the cause of this finding, but the qualitative analysis does suggest some factors.

The second finding, that of a significant effect from the different patients, is also apparent in Table 2, as the pass rate for the different patients varies quite widely. Possibly, PTVs were easier to generate and/or judge for some clinical cases.

A patient/evaluator interaction variable was included in the analysis and was also found to exert a significant effect on the pass rate ($p < 0.05$). The interaction effect would be insignificant only if the rank ordering of each evaluator's scores was rela-

tively consistent in relation to those of the other evaluators across all patients. This was clearly not the case.

With the patient and evaluator variables and interaction accounted for, there was no significant dependence of pass rate upon generator ($p = 0.63$). Therefore, the null hypothesis could not be rejected, i.e., the expert system performance was equally successful to that of human experts.

A qualitative review of the evaluators' reasons for failing PTVs showed them to be classifiable into three categories. PTVs were most frequently rejected for providing too narrow a margin around the tumor volume. A second group was rejected for encroaching upon a critical, radiation sensitive structure. Finally, a few were faulted for overly generous margins. Table 3 lists the frequency of these reasons for the human and computer generators. Evidently, the different generator errors were qualitatively as well as quantitatively similar.

Table 3: Distribution of Reasons for PTV Rejection by Generator

Reason	Human (n=78)	Computer (n=21)
Margin too narrow	69%	62%
Margin too wide	4%	10%
Critical structure	27%	29%

When the qualitative results were analyzed for the different evaluators, a reason for the differences between evaluators begins to become apparent (Table 4).

Table 4: Distribution of Reasons for PTV Rejection by Evaluator

Reason	Eval 1 (n=31)	Eval 2 (n=47)	Eval 3 (n=21)
Margin too narrow	54%	87%	62%
Margin too wide	-	4%	14%
Critical structure	46%	9%	44%

Evaluators 1 and 3 had a similar distribution of reasons for rejecting a PTV. Evaluator 2 differed from the other two markedly. A closer inspection of this evaluator's responses revealed that he carried out an analysis of margins in the Z direction, orthogonal to the CT image plane. Neither of the other evaluators took this step.

DISCUSSION

The results show that the PTV Tool is as successful as human experts in generating a PTV. However, the marked differences in scoring between the different evaluators, and the overall low success rate of the generators, whether human or computer, demonstrate the difficulty of evaluating this system. The results can be seen to relate the high degree of variability in the rating process with the complexity of the output.

One obvious contributor to these discrepancies is the Z-dimension analysis carried out by Evaluator 2. This analysis was appropriate under the evaluators' mutual judging criteria, but apparently it was not addressed separately in their initial discussion, nor did it occur to the other two evaluators. As a consequence of this evaluation, it became clear that the expansion algorithm used in the PTV Tool fails to adequately consider this dimension. (See Figure 2.) However, it also became apparent that the physician generators did not consider this dimension at all. This is an example of an expert system reproducing the errors prevalent in clincal practice. The difficulty of manually producing an accurate three-dimensional expansion from a tomographic image set emphasizes, in itself, the need for an automated tool.

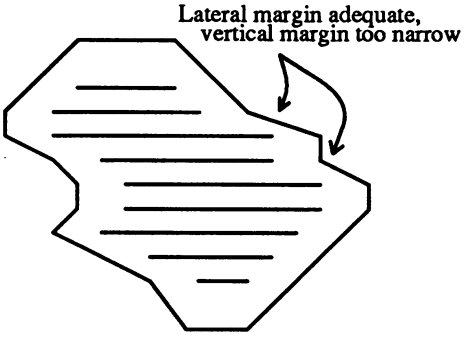

Figure 2: Tumor volume in the x,z plane, of identical shape to Figure 1, showing incorrect expansion discovered at evaluation. CT slice planes are seen on edge in this view.

The failure to consider Z-dimension expansion does not fully explain the results, however. Evaluators 1 and 3 did not consider this factor, yet both delivered relatively low pass rates (40% and 60%) which differ significantly.

These results are, at least in part, attributable to the unusual output of produced by this expert system. In contrast to single-valued expert system output, a geometric solid is inherently multidimen-

sional. Each dimension presents a separate parameter to be considered by the evaluators, who may include additional parameters, as in this study, the relation of the PTV to radiation sensitive critical structures.

The lower scores and decreased evaluator consensus observed in this case likely stem from the large number of parameters upon which the output is judged. When a deviation in any parameter prompts the evaluator to fail the output under review, lower overall scores will be observed as the number of parameters increases, since there are more opportunities for deviation. Furthermore, if differences of opinion or practice exist between evaluators, these will also be exaggerated, as there are more factors over which a difference may be observed.

Expert system output need not be geometric to be multiparametric. In systems producing differential diagnoses as output, inclusion of the reasonable diagnoses in the list and rank ordering of the list represent separate parameters to be considered in evaluation. Frequently, the evaluation of such systems is simplified by creating a heuristic which combines all parameters into a single score.[3, 4] However, when evaluators are presented with the parameters individually, significant variation between evaluators may occur. For example, in their report on the evaluation of ANEMIA, a hematologic expert system, Quaglini, et. al.[5] reported less than 50% consensus among their evaluators.

Just as decrease of evaluator consensus has been observed, scorings lower than one might expect have been reported for systems with complex or multidimensional output. Hickam, et. al.[6] report on the evaluation of ONCOCIN, an expert system designed to recommend treatment for lymphoma patients. Oncologic treatment plans have several variable parameters including treatment type, timing, and dosage, and thus represent a multidimensional output. The ONCOCIN evaluators judged the treatment recommendations of the expert system along with those of the patients' physicians. In cases where these recommendations differed, only 67% of treatment recommendations were judged acceptable, regardless of whether they originated with the human experts or the computer system.

As expert systems continue to advance and proliferate, it is likely that the tasks required of them will increase in complexity such that the number of parameters upon which the outputs will be judged increases. Because of this, evaluations finding lower average scores and significant variability between judges will likely become more common. Consequently, evaluation designs such as the one reported here will be vital to establishing the acceptability of expert system output within the spectrum of clinical practice.

ACKNOWLEDGEMENTS

This work was partially supported by NIH grant no. LM04174 from the National Library of Medicine and contract no. CM97566 from the National Cancer Institute. It is a pleasure to acknowledge the assistance of Jan Halle and Scott Sailer from the University of North Carolina, Joseph Simpson and Clifford Chao from Washington University, St. Louis,, and George Laramore, Keith Stelzer and Tom Griffin from the University of Washington.

References

1. International Commission on Radiation Units and Measurements. *Prescribing, Recording and Reporting Photon Beam Therapy.* International Commission on Radiation Units and Measurements, Bethesda, MD, 1993. Report 50.

2. J. Wyatt and D. Spiegelhalter. Evaluating medical expert systems: What to test and how. *Medical Informatics*, 15(3):205–217, 1990.

3. A. Verdaguer, A. Patak, J. J. Sancho, C. Sierra, and F. Sanz. Validation of the medical expert system PNEUMON-IA. *Computers and Biomedical Research*, 25(6):511–526, December 1992.

4. J. F. Porter, L. C. Kingsland III, D. A. Lindberg, et al. The AI/RHEUM knowledge-based computer consultant system in rheumatology. Performance in the diagnosis of 59 connective tissue disease patients from Japan. *Arthritis and Rheumatism*, 31(2):219–226, February 1988.

5. S. Quaglini, M. Stefanelli, G. Barosi, and A. Berzuini. A performance evaluation of the expert system ANEMIA. *Computers and Biomedical Research*, 21(4):307–323, August 1988.

6. D. H. Hickam, Edward H. Shortliffe, M. B. Bischoff, A. C. Scott, and C. D. Jacobs. The treatment advice of a computer based cancer chemotherapy protocol advisor. *Annals of Internal Medicine*, 103:928–936, 1985.